Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 9th International Conference on Computer Science and Computational Intelligence 2024

A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews

Rio Pramana[a], Marcel Jonathan[a], Habel Steven Yani[a], Rhio Sutoyo[a]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

Emotion recognition within online product reviews is pivotal for enhancing business strategies and customer insights. Recognizing emotions in Indonesian, a language rich in nuances and expressions presents significant challenges due to its complex linguistic structure and the scarcity of tailored datasets. This study aims to bridge this gap by evaluating the effectiveness of Bidirectional Long Short-Term Memory (BiLSTM), Bidirectional Encoder Representations from Transformers (BERT), and ensemble methods in analyzing emotions from Indonesian product reviews using the PRDECT-ID dataset. Extensive fine-tuning across 23 BERT configurations and multiple BiLSTM preprocessing combinations was conducted to adapt these models to the Indonesian linguistic context. Each model was assessed based on the F1 score. The BiLSTM model was particularly effective in configurations with complex preprocessing, achieving an optimal F1 score of 61% through advanced noise removal, stemming, and a modified stopwords list. Conversely, the minimally preprocessed fine-tuned 'base-p2' and 'large-p1' BERT variants achieved F1 scores of 72% and 73%, respectively, both surpassing the previous best result of 71%. This research also explored 4 ensemble methods, combining the strengths of the best-performing BiLSTM and BERT models using both soft-voting and stacked generalization techniques. The unweighted stacked generalization achieved a 74% F1 score, while the weighted method excelled with the highest F1 score of 75%, surpassing other models and highlighting the advantages of strategic model integration. This research significantly advances the development of NLP models for Indonesian text, demonstrating how tailored deep-learning approaches can effectively enhance emotion recognition accuracy.

Peer-review Statement: Peer-review under responsibility of the scientific committee of the 9th International Conference on Computer Science and Computational Intelligence 2024.

Keywords: Deep Learning; Emotion Recognition; Natural Language Processing; Product Review

1. Introduction

Recognizing emotions is crucial in communication, primarily through modern platforms like e-commerce where reviews significantly impact business strategies[1,2]. Natural Language Processing (NLP) is a collection of computa-

* Corresponding author
E-mail address: rsutoyo@binus.edu

Peer-review Statement: Peer-review under responsibility of the scientific committee of the 9th International Conference on Computer Science and Computational Intelligence 2024.

tional methods designed to automatically analyze and describe human language, inspired by theoretical principles[3]. One of its applications is to recognize emotions. In the context of emotion recognition in Indonesian product reviews, several challenges and gaps need to be addressed, such as the complexity of the language structure and the rarity of publicly available Indonesian product review datasets compared to English, which is widely used in other research[4,5]. Therefore, deep learning models for emotion recognition in Indonesian product reviews still require further development and research.

This research evaluates the effectiveness of two advanced deep learning algorithms—Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Encoder Representations from Transformers (BERT)—providing insights into their comparative strengths and the benefits of integrating them through ensemble methods. BiLSTM is chosen for its proficiency in processing sequences from both forward and backward directions. Unlike more traditional models such as Support Vector Machine (SVM) and Naive Bayes, BiLSTM leverages sequential data effectively, making it particularly suited for tasks where the full context of an input sequence is essential for accurate predictions[6]. BERT has been selected due to its robust performance in various NLP tasks and its prevalent use in emotion recognition research[7,8]. Its ability to understand context deeply through pre-trained models makes it highly effective for the nuanced language processing required in emotion recognition. Furthermore, ensemble methods, proven to enhance predictive performance by leveraging the strengths of multiple models, are hypothesized to yield better results in recognizing emotions from Indonesian product reviews[9,10,11]. Based on the research background, here are several research questions (RQ) formulated for this work:

- **RQ1**: In what ways can deep learning models, specifically BiLSTM and BERT, be optimized for accurately recognizing customer emotions in Indonesian language product reviews?
- **RQ2**: What evaluation methods are used to assess the performance of BiLSTM, BERT, and ensemble methods in emotion recognition?
- **RQ3**: How do the BiLSTM algorithm, BERT algorithm, and the application of the ensemble method compare in accurately recognizing emotions in Indonesian language product reviews?

This study introduces advanced deep-learning techniques to the field of emotion recognition in Indonesian product reviews, making the first use of BiLSTM models and ensemble methods for this purpose. This research enhances practical NLP applications by improving the accuracy and robustness of Indonesian emotion recognition models. This research is structured as follows: The literature review section reviews existing theories and prior research relevant to emotion recognition. The methodology section describes the methods and processes employed in developing and evaluating the models. The results and discussion section details the outcomes, analyzing the effectiveness of each model configuration. Finally, the conclusion and future work section summarizes the key findings and suggests pathways for future inquiries into NLP and emotion recognition.

2. Literature Review

2.1. Emotion Recognition Techniques

BiLSTM, an enhanced version of RNN, addresses the issues of derivative explosion and derivative loss in RNN and can handle sequence input of varying lengths, hence mitigating the problem of information loss occurring quickly in circulating neurons[12]. This allows the algorithm to understand better the structure and context of the text being processed. This makes BiLSTM suited for tasks where the full context of an input sequence is essential for accurate predictions[6]. On the other hand, BERT is known for its ability to understand context and relationships between words[4]. BERT uses pre-trained models on large corpora such as Wikipedia or BookCorpus to learn common text representations. Another advantage of the BERT model is its ability to learn data in various languages, including Indonesian, simply by using training resources to adapt the model to the specific language. In addition, BERT uses Masked Language Modeling (MLM) techniques to separate positive and negative text representations related to emotions[13]. This technique allows the model to understand complex and subtle emotions in text. The IndoBERT model is a monolingual BERT language model that has been pre-trained for Indonesian and trained as a Masked Language Model using the Huggingface[14] framework.

2.2. Related Work

Recent studies in emotion recognition have utilized the PRDECT-ID dataset for analyzing Indonesian product reviews. Andres, Jomari Rasheed et al. achieved an F1 score of 71% by applying fine-tuning techniques to the IndoBERT model[7]. Similarly, Hadiwijaya et al. reported an F1 score of 91% using a fine-tuned IndoBERT model for sentiment analysis, demonstrating the effectiveness of transfer learning[8]. Research by Yaquut et al. on classifying emotions from Indonesian tweets via a soft-voting ensemble of SVM and BERT models highlighted an accuracy improvement to 80.23%, surpassing individual model performance[10]. Eggi et al. also utilized a voting ensemble, combining BERT and RoBERTa to classify personalities from social media data, achieving F1 scores of 73% and 74% on two datasets[11]. Feature engineering plays a crucial role in the performance of emotion recognition models. Rhio, et al. found that on average, stemming increases accuracy but reduces F1 score, while data augmentation significantly reduces both accuracy and F1 score. However, their best model, incorporating both stemming and data augmentation with a stratified train-test method, reached an accuracy of 65.27% and an F1 score of 66.09%[15]. This shows that stemming and data augmentation might increase performance depending on the model's configuration. In other research, Al-Omari, H., et al. explored the use of BiLSTM and BERT on English text datasets, where BiLSTM architecture alone achieved a 69% F1 score, and its combination with BERT reached 71%. Their best results came from a weighted ensemble method, producing a top F1 score of 74%[9]. Table 1 summarizes previous research results.

Table 1. The Best *F1* Score Results from Previous Research

Research	Author	Task	Method	Dataset	Dataset Language	F1 Score
[7]	J. R. Andres, et al.	Emotion Recognition	BERT	PRDECT-ID	Indonesian	71%
[8]	M. A. Hadiwijaya, et al.	Sentiment Analysis	BERT	PRDECT-ID	Indonesian	91%
[9]	Al-Omari, H., et al.	Emotion Recognition	Ensemble Method	EmoContext	English	74%
[10]	R. Y. A. A. Ilahi, et al.	Emotion Recognition	Ensemble Method	Twitter	Indonesian	80%
[11]	E. F. Tsani, et al.	Personality Identification	Ensemble Method	Twitter	Indonesian	73%
				YouTube		74%

3. Methodology

This research employs a structured framework, as depicted in Fig. 1. Due to intrinsic differences in these models, there are separate development steps between BiLSTM and BERT. before their evaluation and subsequent integration using ensemble methods.

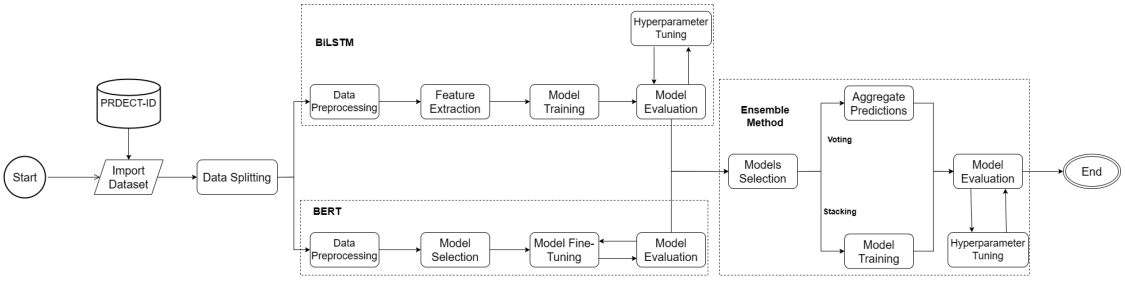

Fig. 1. Research Framework

3.1. Data Collection

The initial step in this research is to collect data obtained from the PRDECT-ID dataset, which contains 5400 Indonesian online product reviews annotated with five emotion labels, validated by a clinical psychology expert. This research uses the PRDECT-ID dataset because it is the only publicly available Indonesian product review dataset annotated with emotions. Furthermore, this dataset's emotion labels were validated and followed criteria set by a clinical psychology expert. This dataset has an unbalanced composition on the number of emotion labels[16], the details of which are illustrated in Fig. 2.

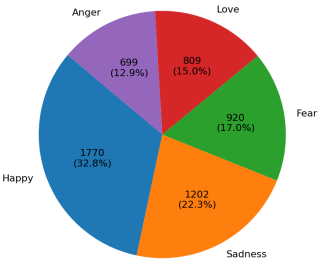

Fig. 2. Emotion Distribution on PRDECT-ID Dataset

3.2. Data Splitting

The dataset was divided using a stratified data splitting method to ensure a balanced representation across training (80%), validation (10%), and testing (10%) sets. This method preserves proportional class representation in each subset, crucial for the dataset's imbalance, and enhances the reliability of the experimental results by ensuring that each subset is representative of the entire dataset[17]. This method thus contributes to the robustness of the research against variability in data distribution. This data split will be used for all model training to provide a fair performance comparison, especially for the ensemble method, as it requires models to be combined and then tested on the same dataset. The next step is data preprocessing, which is separated for BiLSTM and BERT since both require different types of preprocessing.

3.3. BiLSTM

3.3.1. Data Preprocessing

Several preprocessing combinations are applied to prepare the data for BiLSTM model training. The processes included are noise removal, slang translation, stopwords removal, and stemming. The noise removal process first converts all text to lowercase and removes irrelevant information such as numbers and special characters. Slangs present in the text are translated using a combined dictionary from M. S. Saputri, et al. and M. O. Ibrohim, et al. research[18,19]. After that, the stop words that are irrelevant to the emotion recognition task, such as 'yang' and 'ke', are removed. In this research, two versions of the stop words list are used. The original list is from NLTK's Indonesian collection, and the modified version excludes 'tidak', 'baik', and 'kurang' from the list. This research argues that the three removed stopwords are crucial for understanding context in emotion recognition, as these words often carry significant emotional connotations. For example, 'baik', which can be translated as 'good', is generally associated with positive emotions. Meanwhile, 'tidak' and 'kurang' are generally related to negative emotions. Next, words are tokenized and then stemmed using the Sastrawi library, reducing them to their base forms by removing prefixes and suffixes. For example, 'membeli' will be reduced to 'beli'. This process helps standardize variations of the same word, which simplifies the language model and enhances feature extraction efficacy[20]. Not all processes are present in each combination. The specific combinations tested for the BiLSTM model are detailed in Section 4.1.

3.3.2. Feature Extraction

This research uses a fastText pre-trained model for word embedding, specifically cc.id.300.bin, to analyze text data from Indonesian product reviews. FastText is chosen for its robust performance in handling large datasets and its capability to generate word vectors for out-of-vocabulary words through subword information (n-grams), making it particularly suitable for the morphologically rich Indonesian language. It has also shown better performance than Word2Vec and GloVe. Each word in our dataset is represented as a 300-dimensional vector, capturing deep linguistic features essential for effective emotion recognition[21].

3.3.3. Model Training and Hyperparameter Tuning

After complete data preprocessing and feature extraction, this research designed a custom BiLSTM architecture for the model. The pre-trained embeddings are integrated as a non-trainable first layer in our BiLSTM model, ensuring the rich linguistic features are utilized without alteration. The rest of the architecture comprises bidirectional LSTM layers, dropout layers to mitigate overfitting, and dense layers for classification. This research fixed the random seed

at 21 across all training sessions to ensure consistency and reproducibility in our experiments. The model is optimized using the Adam optimizer. The architecture and hyperparameters, such as the number of LSTM units, dropout rates, and learning rates, were iteratively adjusted based on their performance evaluation to refine the model's ability to generalize well on unseen data.

3.4. BERT

3.4.1. Data Preprocessing

Data preprocessing for the BERT model, similar to the BiLSTM approach, includes basic steps like noise removal and slang translation. Given BERT's efficiency with minimal preprocessing, the focus was on employing the least intensive preprocessing necessary to optimize performance[22]. Therefore, this research systematically explored various combinations of noise removal techniques, ranging from minimal to comprehensive approaches. Each combination was methodically tested to determine the optimal set of preprocessing steps that enhance model accuracy and generalization. The specific combinations tested are detailed in Table 2, which illustrates the progression from simple to complete noise removal setups. To ensure proper tokenization, especially in scenarios where special characters are retained, spaces were introduced between words and punctuation marks. For example, the phrase 'bagus,beli' was adjusted to 'bagus, beli' to facilitate correct word separation during tokenization.

Table 2. Noise Removal Type Description

Type	Description
1	Case folding + remove extra whitespaces
2	Case folding + remove extra whitespaces, numbers
3	Case folding + remove extra whitespaces, special characters
4	Case folding + remove extra whitespaces, numbers, special characters

This research also implemented data augmentation methods like synonym replacement, which will generate new review sentences by replacing words with synonyms, thus expanding the diversity of training data without altering the semantic content. For the BERT model, this research primarily used a modified list of stopwords or none at all. This choice was based on the improved performance observed with the modified and unmodified list during the BiLSTM experiments. The exact slang translation and stemming method utilized in the BiLSTM model was employed here. Not all preprocessing techniques were applied in each experiment. For example, while some tests combined stopwords removal and stemming to assess their collective impact, others used only minimal noise removal to examine whether BERT could effectively process text closer to its original form. Given BERT's architecture, which integrates tokenization and embedding generation, complex preprocessing steps are often unnecessary. This allows BERT to efficiently process raw text using its pre-trained, context-rich embeddings, enhancing its ability to capture nuanced linguistic patterns essential for accurate emotion recognition. Detailed descriptions of the specific combinations used are available in Section 4.2.

3.4.2. Model Selection

This research evaluates several variants of the IndoBERT pre-trained models, including IndoBERT-base-p1, IndoBERT-base-p2, IndoBERT-large-p1, and IndoBERT-large-p2. Previous research by Andres, Jomari Rasheed, et al., demonstrated that the IndoBERT-base-p1 model achieved the highest F1-score of 71% on the PRDECT-ID dataset[7]. This research hypothesizes that the larger models, due to their increased number of parameters, might yield better performance. Initially, this research conducted experiments using IndoBERT-base-p1 and IndoBERT-large-p1. Then, the best-performing configurations from these models were tested with IndoBERT-base-p2 and IndoBERT-large-p2, respectively, to determine if newer versions of the models enhance performance.

3.4.3. Model Fine-Tuning

This research undertook the fine-tuning of each model to tailor it specifically to the existing dataset and task. The process is initiated by setting a starting learning rate of 2e-5 and employing a scheduler to adjust this rate across training epochs, enhancing the model's ability to learn effectively without overfitting. The optimization was managed using the Adam optimizer. The gradient clipping technique is incorporated to prevent the gradients from becoming too large and destabilizing the training process[23]. An early stopping mechanism was also implemented to halt training if there were no improvements, effectively saving computational resources and safeguarding the model's performance.

3.5. Ensemble Method

3.5.1. Models Selection

The ensemble method began with selecting the best-performing models from our previous experiments with BiLSTM and BERT. Models were chosen based on their performance metrics on the test dataset. Once the models were selected, this research applied voting and stacked generalization (stacking) to aggregate their predictions.

3.5.2. Voting

Soft voting is employed to aggregate the predictions from the best models. Soft voting takes the probabilities predicted by each model and averages them to produce a final prediction for each review in the test dataset. This approach considers the confidence level of each model's predictions, making it a robust method for ensemble decision-making. In addition, this research also explored a weighted soft voting approach to refine our ensemble method further. The weighted soft voting ensemble function calculates a weighted average of the prediction probabilities, where each model's output is multiplied by a specific weight before averaging. These weights are assigned based on the prior performance of the models, allowing us to leverage the strengths of more accurate models more heavily. This research experimented with multiple combinations of weights to find the best combination. The implementation of both soft and weighted soft voting combines the insights from multiple models, resulting in a new set of probabilities that represent a consensus between the contributing models.

3.5.3. Stacking

This study employed unweighted and weighted methods in stacked generalization (stacking) to effectively combine model outputs. Unweighted stacking involved concatenating prediction probabilities from each contributing model to form a new feature set, which was then used as input for a logistic regression meta-model. This method allowed the meta-model to learn how to integrate predictions from each underlying model based on their collective output on the training data. Weighted stacking involved experimenting with multiple combinations of weights assigned to each model's predictions before concatenation, ensuring all models contributed meaningfully to the final decision. These weighted predictions were then stacked to form the feature set for the logistic regression meta-model. K-Fold cross-validation was used to ensure the model was robust across different subsets of data and generalized well to unseen data, given the modest size of the test dataset of 1080 data.

3.6. Model Evaluation

In the PRDECT-ID dataset, there are indications of imbalance, such as the number of reviews with the emotion of happiness being more significant than other emotions, such as anger. In real-world conditions, happiness and sadness are more accessible and often found in Indonesian product reviews than in the other three emotions [16]. Therefore, a weighted F1 score is chosen to evaluate the model based on its real-world usage. Weighted F1 score is a metric that combines precision and recall for each class. The F1 score for each class will be calculated and then used to calculate the average value of the F1 score, but weight or support will be used. Weighted F1 scores can be used to handle imbalanced datasets while still considering the distribution of each class to describe the real-world situation of the task being carried out. A confusion matrix is added as additional detailed performance information for each model.

4. Results and Discussion

The following is a more detailed explanation regarding the research results of the BiLSTM model, BERT, and the application of the ensemble method.

4.1. BiLSTM

The performance of the BiLSTM model was evaluated under six combinations of preprocessing steps, each increasing in complexity. The results, detailed in Table 3, demonstrate a clear trend: as preprocessing complexity increased, so did the model's F1 score. The most straightforward preprocessing combination resulted in F1 scores of 58% and 60%, while more comprehensive approaches, incorporating slang translation and stemming, improved F1 scores to 59% and 61%. This indicated that while basic preprocessing was implemented, integrating more detailed preprocessing methods enhanced the model's ability to analyze text effectively. The configurations employing a modified stopwords list consistently outperformed those using the original list, with improvements of 2-3% in F1 score. This highlights the value of tailoring the stopwords list to the dataset's specific linguistic traits, optimizing text processing for more

effective emotion recognition. However, an issue was encountered during preprocessing involving the Sastrawi stemming tool, which occasionally led to errors, such as misstemming ”pengemasan” (packing) to ”emas” (gold) instead of ”kemas” (pack). Among the tested configurations, the fourth and sixth combinations achieved the highest F1 scores at 61%. The performance details for these models are illustrated in the confusion matrices in Fig. 3 and Fig. 4.

A qualitative analysis of the confusion matrices reveals that the fourth combination offers better emotion generalization. It minimizes extreme misclassifications, such as misclassifying ’Love’ as ’Anger’, which are significantly different emotional states. Additionally, the fourth combination more effectively classifies under-represented emotions like ’Fear’ and ’Anger’—a notable achievement given their complexity and lower representation in the dataset. While the sixth combination shows a slight advantage in classifying ’Love’, it tends to confuse ’Love’ with ’Happy’, a less significant error given the emotional similarity between the two. Given these observations, the fourth combination is the best model due to its consistent performance across various emotions, particularly in accurately distinguishing between distinctly different feelings. This ability is crucial for applications that require a nuanced understanding of emotions, ensuring the model’s effectiveness in real-world scenarios. Notably, unlike the sixth combination, the fourth combination does not employ slang translation, suggesting that omitting slang translation did not detract from its performance and may indicate limited utility for this preprocessing step in this context.

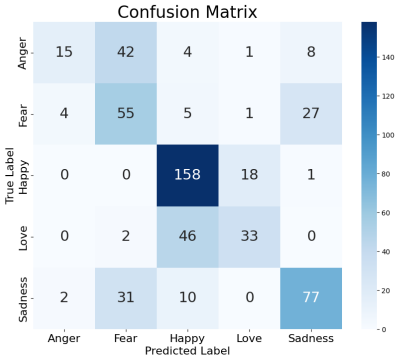

Fig. 3. BiLSTM Combination 4 Confusion Matrix

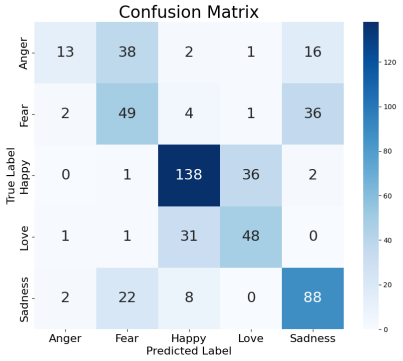

Fig. 4. BiLSTM Combination 6 Confusion Matrix

Table 3. The *F1* Score Comparison of BiLSTM Model

Combination	Noise Removal	Slang Translation	Stopwords Removal	Stemming	F1 Score
1	✓		✓		58%
2	✓		Modified		60%
3	✓		✓	✓	58%
4	✓		**Modified**	✓	**61%**
5	✓	✓	✓	✓	59%
6	✓	✓	Modified	✓	61%

4.2. BERT

This research experimented with a total of 23 configurations of IndoBERT models, with the comprehensive results detailed in Table 4. Analysis of the results revealed a consistent trend where configurations utilizing minimal preprocessing consistently outperformed those with more complex approaches. This was particularly evident in models that only employed type 1 noise removal—the most straightforward form—without any additional modifications such as slang translation, stopword removal, or stemming. These simpler models demonstrated better generalization capabilities as reflected in their F1 scores. The fine-tuned ’base-p2’ model, which utilized synonym replacement for data augmentation while keeping other preprocessing minimal, emerged as the top performer among the base variants with an F1 score of 72%. Similarly, the fine-tuned ’large-p1’ model that applied the most minimal preprocessing achieved the highest F1 score of 73% among the large models. Therefore, both models were selected as the best-performing BERT models. Furthermore, both models also achieved F1 scores that exceeded the best model from the

previous study by Andres, Jomari Rasheed, et al. by 1-2%, demonstrating an improvement in performance using the same dataset and pre-trained IndoBERT model[7]. Given the performance similarities between the base-p2 and large-p1 models (only a 1% difference in F1 score), the base-p2 model might be a preferable choice for practical applications considering its lower computational demands compared to the large variant. This aspect makes the base-p2 model particularly attractive for environments where computational resources are a concern.

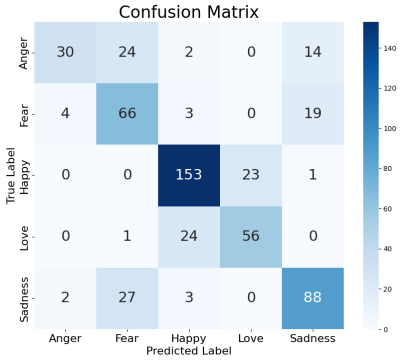

Fig. 5. Best Fine-tuned IndoBERT base-p2 Confusion Matrix

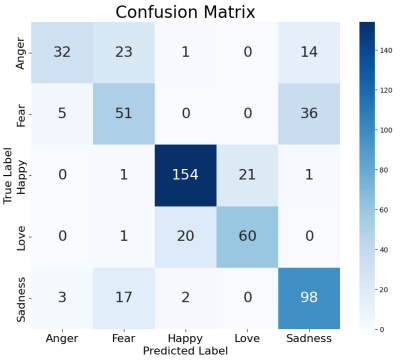

Fig. 6. Best Fine-tuned IndoBERT large-p1 Confusion Matrix

This performance is also consistent with the observed trend that larger models tend to perform better than base models under identical configurations. These results emphasized that less intervention can indeed be more effective in specific natural language processing applications, especially in the context of advanced pre-trained models like IndoBERT. The confusion matrices for the selected BERT models are shown in Fig. 5 and Fig. 6. Both models demonstrate a significant improvement in correctly identifying 'Anger' compared to the BiLSTM model. The fine-tuned base-p2 model excels in recognizing 'Fear', although it occasionally misclassifies 'Sadness' as 'Fear'. On the contrary, the fine-tuned large-p1 model shows superior performance in accurately identifying 'Sadness', but it tends to confuse 'Fear' with 'Sadness'. These matrices suggest that while both models perform well in recognizing distinct emotional expressions and show improvement with 'Anger', they still struggle with 'Fear' and 'Sadness', often confusing these emotions with each other.

4.3. Ensemble Method

The ensemble methods combined the best performing fine-tuned 'base-p2' and 'large-p1' IndoBERT models with the BiLSTM model in 4 distinct configurations. Table 5 shows the results of ensemble method experiments using four different configurations. The unweighted soft-voting ensemble achieved an F1 score of 72%, matching the performance of the fine-tuned 'base-p2' model. The weighted soft-voting ensemble improved upon this, reaching an F1 score of 73%. The best weights that combine all three models are 0.15 for BiLSTM, 0.3 for fine-tuned base-p2, and 0.55 for fine-tuned large-p1. The weights reflect the relative confidence in each model's predictive accuracy, with the highest weight given to the fine-tuned 'large-p1' model due to its superior individual performance. In contrast to the soft-voting methods, which could only match the performance of the best individual model in the ensemble, the stacked generalization approaches surpassed all models in the ensemble, demonstrating their superior effectiveness. The unweighted stacked generalization method surpassed all individual models in the ensemble by achieving a 74% F1 score. Moreover, the weighted stacked generalization which optimally adjusted the contributions of each model with weights of 0.15 for BiLSTM, 0.35 for fine-tuned base-p2, and 0.5 for fine-tuned large-p1, achieved the highest F1 score of 75%, outperforming all other models. These results highlight the considerable benefits of employing weighted ensemble methods, especially stacked generalization, for complex tasks like emotion recognition. The model that achieved the highest F1 score of 75% sets a new standard by demonstrating how effectively diverse learning algorithms can be combined to improve prediction accuracy. This approach significantly enhances overall performance by aligning the ensemble's output with the most reliable predictions from its constituent models.

Table 4. The *F1* Score Comparison of Fine-tuned IndoBERT Model

Pre-Trained Model	Noise Removal	Slang Translation	Stopwords Removal	Stemming	Data Balancing	*F1* Score
Base-p1	Type 1					71%
Base-p1	Type 1	✓				70%
Base-p1	Type 1	✓	Modified			69%
Base-p1	Type 1	✓	Modified	✓		66%
Base-p1	Type 1				Class Weighting	70%
Base-p1	Type 1				Synonym Replacement	71%
Base-p1	Type 2	✓	Modified	✓		67%
Base-p1	Type 4	✓	Modified	✓		67%
Base-p1	Type 4					68%
Base-p2	Type 1					71%
Base-p2	**Type 1**				**Synonym Replacement**	**72%**
Large-p1	**Type 1**					**73%**
Large-p1	Type 1	✓				71%
Large-p1	Type 1	✓	Modified			69%
Large-p1	Type 1	✓	Modified	✓		69%
Large-p1	Type 1				Class Weighting	71%
Large-p1	Type 1				Synonym Replacement	72%
Large-p1	Type 2	✓	Modified	✓		65%
Large-p1	Type 3					71%
Large-p1	Type 3	✓				71%
Large-p1	Type 4	✓	Modified	✓		67%
Large-p1	Type 4					70%
Large-p2	Type 1					72%

Table 5. The *F1* Score Comparison of Ensemble Method

Method	Model 1		Model 2		Model 3		F1 Score
Soft-voting	BiLSTM		IndoBERT Base-p2		IndoBERT Large-p1		72%
Stacked Generalization	BiLSTM		IndoBERT Base-p2		IndoBERT Large-p1		74%
Method	**Model 1**	**Weight 1**	**Model 2**	**Weight 2**	**Model 3**	**Weight 3**	**F1 Score**
Soft-voting (Weighted)	BiLSTM	0.15	IndoBERT Base-p2	0.3	IndoBERT Large-p1	0.55	73%
Stacked Generalization (Weighted)	**BiLSTM**	**0.15**	**IndoBERT Base-p2**	**0.35**	**IndoBERT Large-p1**	**0.5**	**75%**

5. Conclusions and Future Works

This research evaluated the effectiveness of BiLSTM, BERT, and ensemble methods for emotion recognition in Indonesian online product reviews using the PRDECT-ID dataset. Our findings demonstrate that each method has unique strengths and contributes valuable insights into emotion recognition. The BiLSTM model, tested under multiple preprocessing configurations, shows that increasing the complexity of the preprocessing steps consistently improved its performance metrics. The best-performing BiLSTM configuration achieved an F1 score of 61% by utilizing comprehensive noise removal, stemming, and a modified stopwords list, which indicates the importance of tailoring preprocessing techniques to the linguistic characteristics of the dataset. This research thoroughly tested 23 configurations of BERT involving four distinct pre-trained models. The models, specifically the fine-tuned 'base-p2' and 'large-p1' with the most minimal preprocessing, stood out in our experiments. The 'base-p2' model achieved an F1 score of 72%, while the 'large-p1' achieved the best F1 score of 73%, validating the hypothesis that minimal preprocessing can leverage BERT models more effectively and larger models with higher computational cost would perform better. This performance surpassed the earlier best model's F1 score by 1-2% and highlighted the benefits of our fine-tuning strategy, which fine-tunes the same model and dataset used in previous research. The ensemble approaches combined the strengths of the best individual BiLSTM and BERT models. Four different configurations were tested, utilizing both soft-voting and stacked generalization methods. The unweighted stacked generalization achieved a 74% F1 score, while the most successful ensemble method was the weighted stacked generalization, which achieved the highest F1 score of 75%. This approach surpassed the performance of all individual models and optimized the output by adjusting the weights based on the reliability of each model's predictions. These findings highlight the potential of ensemble methods in improving the robustness and accuracy of emotion recognition systems. Future work should focus on refining preprocessing techniques to better capture the nuances of the Indonesian language, improving the

accuracy of stemming, expanding the dataset for broader emotional coverage, and optimizing models through targeted hyperparameter tuning. These steps will improve the models' performance and pave the way for more effective NLP applications in Indonesian emotion recognition.

Acknowledgement

This work is supported by Bina Nusantara University as a part of Bina Nusantara University's BINUS International Research - Applied entitled "Towards Virtual Humans Framework with Emotions Model as Indonesian Folklore Storyteller" with contract number: 069C/VRRTT/III/2024 and contract date: March 18, 2024.

References

1. Syahputra, H.. Sentiment analysis of community opinion on online store in indonesia on twitter using support vector machine algorithm (svm). In: *Journal of Physics: Conference Series*; vol. 1819. IOP Publishing; 2021, p. 012030.
2. Faizal, B., Abraham, S.. Nlp based automated business report summarization. In: *2022 International Conference on Innovative Trends in Information Technology (ICITIIT)*. IEEE; 2022, p. 1–4.
3. Chowdhary, K.. *Fundamentals of artificial intelligence*. Springer; 2020.
4. Lek, J.X.Y., Teo, J., et al. Academic emotion classification using fer: A systematic review. *Human Behavior and Emerging Technologies* 2023;**2023**.
5. Chowanda, A., Sutoyo, R., Tanachutiwat, S., et al. Exploring text-based emotions recognition machine learning techniques on social media conversation. *Procedia Computer Science* 2021;**179**:821–828.
6. Caterini, A.L., Chang, D.E.. *Recurrent Neural Networks*. Cham: Springer International Publishing. ISBN 978-3-319-75304-1; 2018, p. 59–79. doi:\bibinfo{doi}{10.1007/978-3-319-75304-1_5}. URL https://doi.org/10.1007/978-3-319-75304-1_5.
7. Andres, J.R., Soetandar, J.P., Sutoyo, R., Riza, H., et al. Emotion recognition model using product review from indonesia marketplace. In: *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*. IEEE; 2023, p. 67–71.
8. Hadiwijaya, M.A., Pirdaus, F.P., Andrews, D., Achmad, S., Sutoyo, R.. Sentiment analysis on tokopedia product reviews using natural language processing. In: *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*. IEEE; 2023, p. 380–386.
9. Al-Omari, H., Abdullah, M.A., Shaikh, S.. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In: *2020 11th International Conference on Information and Communication Systems (ICICS)*. 2020, p. 226–232. doi:\bibinfo{doi}{10.1109/ICICS49469.2020.239539}.
10. Ilahi, R.Y.A.A., Derwin, S.. Emotion classification of indonesian twitter social media text using soft voting ensemble method. *-B:* 2024; **15**(01):101.
11. Tsani, E.F., Suhartono, D.. Personality identification from social media using ensemble bert and roberta. *Informatica* 2023;**47**(4).
12. Cai, R., Qin, B., Chen, Y., Zhang, L., Yang, R., Chen, S., et al. Sentiment analysis about investors and consumers in energy market based on bert-bilstm. *IEEE access* 2020;**8**:171408–171415.
13. Hoang, M., Bihorac, O.A., Rouces, J.. Aspect-based sentiment analysis using bert. In: *Proceedings of the 22nd nordic conference on computational linguistics*. 2019, p. 187–196.
14. Wilie, B., Vincentio, K., Winata, G.I., Cahyawijaya, S., Li, X., Lim, Z.Y., et al. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 2020, .
15. Sutoyo, R., Spits Warnars, H.L.H., Isa, S., Budiharto, W.. Indonesian twitter emotion recognition model using feature engineering. *International Journal of Advanced Computer Science and Applications* 2023;**14**:1057–1065. doi:\bibinfo{doi}{10.14569/IJACSA.2023.01412108}.
16. Sutoyo, R., Achmad, S., Chowanda, A., Andangsari, E.W., Isa, S.M.. Prdect-id: Indonesian product reviews dataset for emotions classification tasks. *Data in Brief* 2022;**44**:108554.
17. Sadaiyandi, J., Arumugam, P., Sangaiah, A.K., Zhang, C.. Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset. *Electronics* 2023;**12**(21):4423.
18. Saputri, M.S., Mahendra, R., Adriani, M.. Emotion classification on indonesian twitter dataset. In: *2018 International Conference on Asian Language Processing (IALP)*. 2018, p. 90–95. doi:\bibinfo{doi}{10.1109/IALP.2018.8629262}.
19. Ibrohim, M.O., Budi, I.. Multi-label hate speech and abusive language detection in Indonesian Twitter. In: Roberts, S.T., Tetreault, J., Prabhakaran, V., Waseem, Z., editors. *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics; 2019, p. 46–57. doi:\bibinfo{doi}{10.18653/v1/W19-3506}. URL https://aclanthology.org/W19-3506.
20. Pramana, R., Subroto, J.J., Gunawan, A.A.S., et al. Systematic literature review of stemming and lemmatization performance for sentence similarity. In: *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*. IEEE; 2022, p. 1–6.
21. Dharma, E.M., Gaol, F.L., Warnars, H., Soewito, B.. The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. *J Theor Appl Inf Technol* 2022;**100**(2):349–359.
22. Kurniasih, A., Manik, L.P.. On the role of text preprocessing in bert embedding-based dnns for classifying informal texts. *Neuron* 2022; **1024**(512):256.
23. Acharya, V., Dhiman, G., Prakasha, K., Bahadur, P., Choraria, A., Sushobhitha, M., et al. Ai-assisted tuberculosis detection and classification from chest x-rays using a deep learning normalization-free network model. *Computational Intelligence and Neuroscience* 2022;**2022**.